Oligo DNA chip을 이용한 급성 골수성 백혈병의 분류와 항암제 적합성 예측 시스템 개발에 관한 연구

공현중※, 김희찬, 박한수[1], 서정선[1], 박숙련[2], 박선양[2]
서울대학교 대학원 협동과정 의용생체공학전공※,, 의과대학 의학공학교실 및 의학연구원
의용생체공학연구소, 생화학교실[1], 내과학교실[2]

Research about oligo DNA microarray analysis system for the classification of acute myeloid leukemia and the prediction of anti-cancer drug response

H. J. Kong※, H. C. Kim, H. S. Park[1], J. S. Seo[1], S. R. Park[2], S. Y. Park[2]
Interdisciplinary Program, Medical and Biological Engineering Major, Graduate School※„
and Dept. of Biomedical Engineering, and Biochemistry[1], and Internal Medicine[2], College of
Medicine. Seoul National Universitv

ABSTRACT

In this research to minimize side effects of anti-cancer treatment to AML patients, we developed the system that can predict responses for the conventional anti-cancer drug dosage using a new print type oligo DNA chip. After pre-processing(gene filtering, normalization and missing value estimation) seven AML patients' expression data, we can extract 89 genes that is significant to classify the CR(complete remedy) group and NoCR(no complete remedy) group using SAM method. And through applying hierarchical clustering to that data, two CR samples and 5 NoCR samples can be classified. If we make a new biochip with significant genes, effective anti-cancer treatments will be available.

Introduction

암은 전체 사망 중 가장 큰 원인 질환으로, 인류의 세 명중 한 명에서 암이 발생하고 있고, 국내에서도 갈수록 큰 증가세를 보이고 있다. 20세기 이후 현대 의학은 암에 대한 연구와 치료법을 발전시켜왔는데, 화학요법에 민감하게 반응하는 암에 관심이 모아져 왔다. 항암 화학요법에서는 임상 연구에서 통계적으로 효과적인 항암제 투여 후, 효과가 없으면 그 다음으로 효과적인 항암제를 투여하는 방법을 사용해왔는데, 환자의 입장에서 그 특정 항암제가 본인에게 효과가 있는지의 여부를 모른 상태에서 골수 억제, 신독성, 신경독성 등의 부작용을 감수해야만 했다.

최근 인간유전체사업(HGP)의 완성 이후, 각종 high throughput 도구들에 의해 쏟아져 나오는 대량의 genome, proteome 데이터를 각종 IT 기술을 이용해 유용한 지식으로 가공하는 일이 중요한 이슈로 떠오르게 되었는데, 그중에 DNA Chip (microarray)은 생체 내 발현 유전자의 활성을 동시에 대규모로 모니터링을 가능케 해줌으로써 현재 기능유전체의학 연구에서 핵심 기술이 되고 있다. 하지만 기존에 많이 쓰이고 있는 cDNA chip과 Affymetrix사의 oligo DNA chip은 국내에서 연구를 하기에는 많은 제약 사항을 지니고 있었다. 본 연구에서는 이들이 지니고 있는 제약 사항을 극복할 수 있는 새로운 개념의 print type의 oligo DNA chip[1]을 사용하여, 항암제로 효과를 보이는 암 중 하나인 급성 백혈병, 그중에서도 급성골수성백혈병(AML) 환자를 대상으로 개인 차이에 따른 항암제의 부작용을 극소화할 수 있도록 항암제 적합성 예측을 위한 분석시스템을 개발하였고, 최종적으로 항암제 적합성 예측 바이오칩 제작을 목표로 하고 있다.

Methods

Datasets & Experiment Design

서울대 병원 종양 내과에서 급성 골수성 백혈병 진단으로 내원한 환자들의 골수를 생검하여 암세포 부분이 있는 mononuclear층에서 total RNA를 뽑고, 또한 control로 사용할 일반인의 Jurkat T lymphocyte로부터 mRNA를 추출한 후, 역전사를 통해 각각에 대한 cDNA를 얻었다. 이 둘 각각에 Cy3(green)와 Cy5(red) dye를 부착시킨 후, 각종 질병들과 관계된 유전자들을 포함하는 10416개의 유전자들의 oligomer(50bp) DNA를 spotting시킨 oligo DNA chip[1]위에 둘을 섞어 뿌려줌으로써, 발현된 유전자들이 Chip에 상보적으로 결합하도록 하였다. 그리고 각 환자들에게 기존의 표준 프로토콜로 항암제 투여 후 완전 치료 여부를 확인하였는데, 이는 개발된 예측 시스템이 올바르게 적합성 여부를 가려냈는지 판별하고, supervised learning 기법을 사용할 경우, 새로운 환자 샘플에 대해 항암제 적합성을 예측하기 위해 필요하다.

Pre-processing

스캐너에서 녹색과 빨강 레이저를 각각 chip에 쪼여 hybridization된 spot에서 내는 형광을 읽어 들여 각각에 대해 세기를 수치화하였다(각각을 편의상 G와 R로 나타냄). 이 과정에서 poor spot이나 empty spot은 NA라고 missing value로 처리했다. 그림1은 스캐너로 읽은 spot 세기 값을 바탕으로 pseudo coloring해서 얻은 dna chip에서의 유전자의 발현의 모습이다.

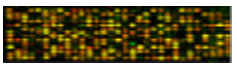

그림1

본격적인 분석을 위해 위에서 얻은 R과 G값을 그대로 사용하지 않고, 감광도의 순서를 유지하면서 작은 차이를 강조해 주며, 부수적으로 정규분포에 가까워지게 해주기 위해 R과 G에 로그를 취한 후, 그것의 기울기 축을 중심으로 데이터를 바로 보게 함으로써 잘못 측정된 유전자를 시각적으로 쉽게 구분해주는 효과를 주기 위해, cy3와 cy5 채널의 세기 값의 비율(M)과 평균(A) 형태로 변화해 주었다(그림2(b) 참조)[2]. 그리고 전체 data(10416x7개) 중 약 20% 차지하고 있는 NA값에 대해, 10416개의 gene 각각에 대해 발현값이 7개의 chip들에서 각각 실수가 아닌 NA로 처리된 게 있으면 그 개수의 비율이 80%를 상회할 경우 그 gene은 없애는 gene filtering을 거친 후 10416개의 gene에서 6483개의 gene이 남았다.

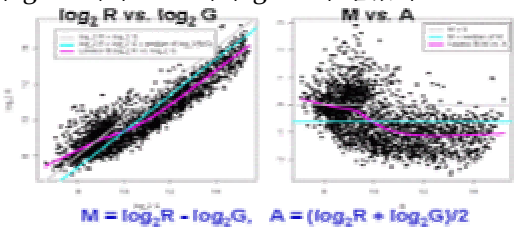

그림2(a), 그림2(b)

그런데 측정된 형광 세기에는 여러 가지 환경적 요인들(e.g. different labeling efficiencies of the dyes, different amounts of Cy3- and Cy5-labelled cDNA, different scanning parameters, print-tip, spatial or plate effects 등)에 의한 systematic variation이 존재하기 때문에 이러한 영향을 제거하기 위해 normalization이라는 과정이 필수적 그 대상은 chip들과 한 chip 안에 존재하는 print-tip group들이다. 한 print-tip group 내의 M vs. A plot에서 local regression 방법 중 하나인 LOWESS[3]를 이용해 fitting line을 구해 원래 M값에서 빼줌으로써 A값에 따라 다르게 M의 위치를 상하로 조정해줌으로써 location normalization을 수행한 후, MAD를 이용해 한 chip안에 존재하는 24개의 print-tip group들 간에 M값의 분산이 같도록 해주는 scale normalization을 수행하였다[2]. 마찬가지 방법으로 나머지 6장의 chip들 각각에 대해 수행한 후(within-slide normalization), 서로 다른 분산을 갖고 있는 7장의 chip의 M값들에 대해 앞서와 같은 scale normalization을 함으로써 분산이 같도록 만들어주었다(between-slide normalization). 그림3에서 (a)는 normalization 전, (b)는 location normalization 후, (c)는 scale normalization 후의 효과를 나타내주는, M값에 대한 box plot이다.

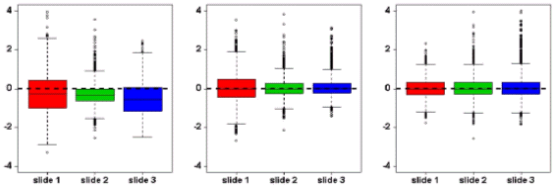

그림3(a) 그림3(b) 그림3(c)

앞서 언급한 gene filtering을 거쳤어도 6704 by 8의 M값 matrix 상에는 여전히 NA로 처리된 값들이 존재한다. missing value에 대한 estimation은 robustness가 떨어지는 clustering algorithm을 위해서는 필수적인데, 본 연구에서는 KNN (k-nearest neighborhood)을 이용해 NA 값을 추정하였다[4].

Clustering(Gene selection & Classification)

전처리가 된 총 6483개의 gene들에서 급성 골수성 백혈병에서 항암제 적합성 여부를 구분시켜주는데 중요한 marker gene들을 추출해야 하는데, 단순히 m-fold change가 이루어진 gene들을 선택하는 것은 A값이 작은 gene에 대해 M값이 과장될 수 있는 등의 이유로 위험하기 때문에, 본 연구에서는 t-test의 변형이며 비모수적 방법인 SAM[5]이라는 기법을 사용하였다. 간단히 설명하자면 SAM은 모든 유전자 발현량의 d-statistics 통계량을 permuted d-statistics 통계량과 비교하여 order statistics를 얻음으로써, 한 개의 threshold 값(delta)으로 유의수준을 결정하고, 우연히 중요한 유전자라고 판단할 수 있는 가능성(false discovery rate)을 알려준다. 이렇게 해서 89개의 marker 유전자들을 찾았다.

이렇게 얻은 89(gene개수) by 7(sample개수)의 M값 matrix에 대해 통계학에서 고전적인 clustering 방법 중 하나인 hierarchical clustering[6]을 적용하였는데, 7개의 sample vector들 간에 상관계수로 유사도 행렬을 만든 후, sample들 간의 유사도가 큰 것끼리 순서대로 서로 묶어 나가도록 하는 게 이 기법의 핵심이다. 이를 이용하여 7개의 AML sample들을 항암제 적합성이 있는 그룹(5개)과 아닌 그룹(2개)의 두 가지 class로 구분할 수 있었다. 관점을 바꿔 gene vector에 대해서도 마찬가지 방법을 수행하면 89개의 gene들 간의 연관 관계를 볼 수 있었다. 그림4에서 세로는 sample, 가로는 gene들 간의 clustering 결과이다.

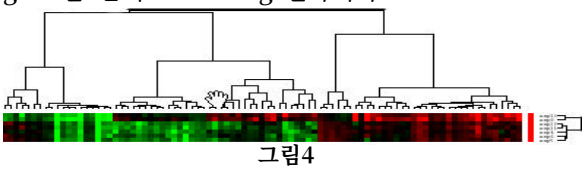

그림4

Discussion

지금까지 항암 화학 요법은 개인별 효과나 부작용의 차이가 명백히 있으나, 현재까지의 연구를 통해 통계적으로 효과가 있는 항암제를 일률적으로 투여할 수밖에 없었다. 많은 샘플을 확보하고 계속적인 연구를 통해 확립된 유전자들만으로 바이오칩을 만들어 진단에 사용한다면 적은 비용으로 개개인에 대한 항암제의 민감성을 알거나, 일정 항암제에 대한 개개인의 부작용을 예측할 수 있어 환자와 의사 모두에게 큰 도움이 되리라 생각한다.

Reference

[1] http://http://www.macrogen.com

[2] Yee Hwa Yang et all. Normalization for cDNA miroarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 2002, vol. 30, No. 4 e15

[3] Cleveland, W.S., *Robust locally weighted regression and smoothing scatterplots.* Journal of the American Statistical Association, 74:829-835, 1979.

[4] Troyanskaya, O., et al.,*Missing value estimation methods for DNA microarrays*. Bioinformatics, Vol. 17 no. 6 2001, pages 520-525.

[5] Tibshrani, Robert et al. *Significance analysis of microarrays applied to the ionizing radiation response.* PNAS 2001 98: 5116-5121, (Apr 24).

[6] Johnson, Richard A., Applied multivariate statistical analysis. Prentice Hall Internatinal Edition, pages